

# Genetic Heterogeneity Estimated by RAPD Polymorphism of Four Tuber-bearing Potato Species Differing by Breeding System

J. B. Bamberg<sup>1\*</sup> and A. H. del Rio<sup>2</sup>

<sup>1</sup>USDA/Agricultural Research Service, Vegetable Crops Research Unit, Inter-Regional Potato Introduction Station, 4312 Hwy. 42, Sturgeon Bay, WI 54235, USA

<sup>2</sup>Department of Horticulture, University of Wisconsin-Madison, 1575 Linden Drive, Madison, WI, 53706, USA

\*Corresponding author: Tel: 920-743-5406; Fax: 920-743-1080; Email: nr6jb@ars-grin.gov

## ABSTRACT

Most wild potato germplasm in genebanks is collected, preserved, and evaluated as botanical seed populations that may be highly heterozygous and heterogeneous. However, some species are selfers so potentially very homozygous, perhaps also homogeneous. Intrapopulation heterogeneity increases sampling error that can undermine consistency in seed regeneration in the genebank, screening results, germplasm collecting, and estimates of taxonomic relationships. Thus, knowledge of genetic heterogeneity (GH) can predict the need to commit more resources for larger sample sizes or replication when populations of a given species are being regenerated, evaluated, collected, and classified. This study investigated within-population GH in 32 potato populations comprising four different breeding systems observed in *Solanum* species: *S. fendleri* (2n=4x=48, disomic selfer), *S. jamesii* (2n=2x=24, outcrosser), *S. suurense* (2n=4x=48, tetrasomic outcrosser), and *S. verrucosum* (2n=2x=24, selfer). RAPD markers were used to estimate heterogeneity among 24 individuals per population. Populations of *S. verrucosum* were quite homogeneous with an average GH of 6.0%. Similarly low heterogeneity was detected among the eight populations of *S. fendleri* (average GH=7.1%). In contrast, *S. jamesii* and *S. suurense* had a much higher GH of 29.4% and 44.1%, respectively. These results demonstrate and quantify the great difference in intrapopulation heterogeneity among wild potato species. Calculations based on

intrapopulation heterogeneity indicate that samples should be composed of 25 to 30 random plants for low sample variation that is uniform for all species.

## RESUMEN

La mayor parte de germoplasma de papa silvestre existente en los bancos de genes es recolectada, preservada y evaluada como poblaciones de semilla botánica, las mismas que pueden ser sumamente heterocigotas y heterogéneas. Sin embargo, algunas especies son autóгамas, y de este modo potencialmente homocigotas y puede ser que sean también homogéneas. La heterogeneidad intrapoblacional, aumenta el error de muestreo, y puede atentar contra varios aspectos entre ellos: la consistencia de regeneración de la semilla en el banco de genes, los resultados del tamizado, la recolección de germoplasma y los estimados de la relación taxonómica. De este modo, el conocimiento de la heterogeneidad genética (GH) puede predecir la necesidad de mayores recursos para un tamaño dado de muestra o réplica, cuando la población de una especie determinada está siendo regenerada, evaluada, recolectada y clasificada. Este estudio investigó dentro de la población GH, 32 poblaciones de papa en cuatro sistemas diferentes de mejoramiento, observados en especies de *Solanum*: *S. fendleri* (2n =4x =48, disómico autóгамo), *S. jamesii* (2n= 2x=24, de polinización cruzada), *S. suurense* (2n=

**4x=48, tetrasómico de polinización cruzada) y *S. verrucosum* (2n=2x=24, autógeno). Para estimar la heterogeneidad entre 24 individuos por población se utilizaron marcadores RAPD. Las poblaciones de *S. verrucosum* fueron totalmente homogéneas con un promedio GH de 6.0%. De la misma manera se detectó una baja heterogeneidad entre ocho poblaciones de *S. fendleri* (promedio GH = 7.1%). En cambio *S. jamesii* y *S. sucrense* tuvieron un GH mucho más alto, 29.4% y 44.1% respectivamente. Estos resultados demuestran y cuantifican la gran diferencia que existe en heterogeneidad intrapoblacional entre las especies silvestres de papa. Cálculos basados en la heterogeneidad de una intrapoblación indican que las muestras deben estar compuestas de 20-25 plantas tomadas al azar para mantener la variación que es uniforme para todas las especies.**

## INTRODUCTION

The US Potato Genebank (NRSP-6) preserves accessions of more than 140 potato species (Bamberg et al. 1996). These include a range of different breeding systems and ploidy levels (Correll 1962; Hawkes 1990). Most wild potato species are collected, propagated, preserved, and evaluated as botanical seed, so individuals within outcrossing populations could be highly heterozygous and heterogeneous. However, some selfing species would tend toward homozygous individuals, perhaps also homogeneous populations.

The genetic structure of a population indicates the allelic diversity it contains and dictates the best strategies for germplasm conservation and enhancement. For instance, a hypothetical completely homogeneous inbred population has the minimum number of alleles per locus (1), so is the most efficient subject for germplasm characterization and evaluation, since any single plant is a completely representative sample of the genetics of the population. Such populations are also at the lowest risk for loss of diversity during sexual reproduction, so can be efficiently multiplied using fewer resources than heterogeneous populations that contain segregating alleles that may be vulnerable to loss (Bamberg and del Rio 2003; del Rio and Bamberg 2003).

RAPDs provide many markers with which to characterize genetic structure (Lynch and Milligan 1994), and some previous studies have applied these and other markers to study variation within potato germplasm populations. Our previous

research identified *Solanum jamesii* (*jam*) as a highly heterogeneous species (del Rio et al. 1997b), and we also used RAPDs to assess the homogeneity of populations of *S. polytrichon* and *S. stoloniferum* with high insect resistance as an indication of the need for "fine" (genotype-level) screening for the selection of parents for breeding (Bamberg et al. 2000). Hosaka and Hanneman (1991) used seed protein variation to identify *S. verrucosum* (*ver*) as one of the most homogeneous of numerous potato species they tested. Spooner et al. (1992) used isozymes to show that individuals within populations of the highly inbreeding species of series *Etuberosa* are also very uniform. Oliver and Zapater (1984) used isozymes to show high heterozygosity and heterogeneity within cultivated tetrasomic tetraploid potato species.

This investigation empirically measured within-population genetic heterogeneity among potato species representing four different breeding systems: self-incompatible outcrosser *jam*, self-compatible diploid *ver*, self-compatible disomic polyploid *S. fendleri* (*fen*), and tetrasomic tetraploid *S. sucrense* (*scr*) with RAPDs. The objective was to measure GH and its variability among species, and to examine the potential impact on potato germplasm conservation and use.

## MATERIALS AND METHODS

Seven populations of *jam*, eight populations each of *fen* and *scr*, and nine of *ver* were used (Table 1). RAPD markers (number) polymorphic within each species were generated: *jam* (48), *ver* (41), *fen* (58), *scr* (35), by the method described in del Rio et al. (1997a). Twenty-four individuals from each population were used. RAPD loci were classified according to the percent of banded plants (BP) in 10% increments: exactly 0%, rounding to 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, or exactly 100%. Genetic heterogeneity (GH) of a population was measured by averaging the theoretical proportion of heterozygous individuals at each locus (considering only loci polymorphic within species).

GH was calculated from the bandless allele frequencies assuming unlinked diallelic RAPD loci. For example, for diploids, consider the two possible alleles for any given RAPD locus: (B) the band, and (b) the blank, having frequencies *B* and *b*, respectively. Now *b* is the square root of the blank-plant (bb) frequency, and *B* is 1-*b*. All plants not homozygous (BB) or (bb) are heterozygotes. Thus GH, the frequency of heterozygotes is 1-(*b*<sup>2</sup> + *B*<sup>2</sup>). This calculation, which assumes equi-

TABLE 1—*Estimated genetic heterogeneity, GH (proportion of heterozygous individuals).*

<i>S. fendleri</i>		<i>S. jamesii</i>		<i>S. sucrense</i>		<i>S. verrucosum</i>	
PI	GH	PI	GH	PI	GH	PI	GH
275157	13.1*	275169	36.8*	473506	38.0*	161173	9.0*
275160	6.3	458423	29.9	473532	33.6*	275256	2.1*
275161	5.1*	458424	26.6	498286	47.4	275258	7.4
275163	6.9	458425	27.1	498300	49.3	275260	7.9
564028	6.6	458426	32.0	498301	55.8*	310966	5.6
564031	10.5*	458427	32.0	498302	49.3*	365404	3.0
564039	3.9*	564074	21.6*	498306	43.5	558483	4.9
564040	4.1*			6798	35.8	558485	7.2
						570643	6.5
average	7.1		29.4		44.1		6.0
95% confidence interval	5.2-9.1		25.7-33.3		39.4-48.8		4.0-8.2

GH values with "\*" are outside of the 95% confidence interval for the species mean.

librium, is valid regardless of the species' reputed breeding system in nature, since all populations at the US Potato Genebank (including those tested here) have been sexually propagated by at least two generations of random intermating.

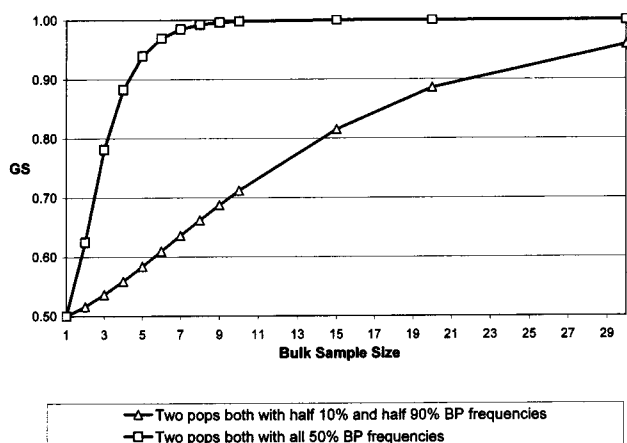
Heterogeneity's important impact is sample variation. If one knows the BP frequency, random bulk sample variation can be calculated. An easily grasped standard for measuring the effects of sample variation is to calculate the apparent similarity of random duplicate samples from a single population, i.e., a comparison for which the true genetic similarity (GS) is 100%. If duplicate samples from a single population have a relatively low GS, resolution of the true relationships of different populations will be difficult. This approach is completely analogous to the measurement and control (minimization) of experimental error in basic parametric statistics in order to quantify the significance of the differences between means. We therefore asked the question: What would be the *apparent* average GS of an infinite number of random pairs of samples composed of different sized bulks taken from populations of the same species (actual GS = 100%)?

To answer this question we used the BP frequency at each locus of each population to calculate the expected frequency of banded samples when bulking DNA from one to 30 random plants. From this we calculated the probability of random samples matching at each locus (i.e., the apparent GS for that single locus), averaged the expected GS of all loci within a species, and graphed against bulk sample size. This can be illustrated with familiar probability problems involving the flipping of coins. Let all coins be identical, with heads (H) and tails (T)

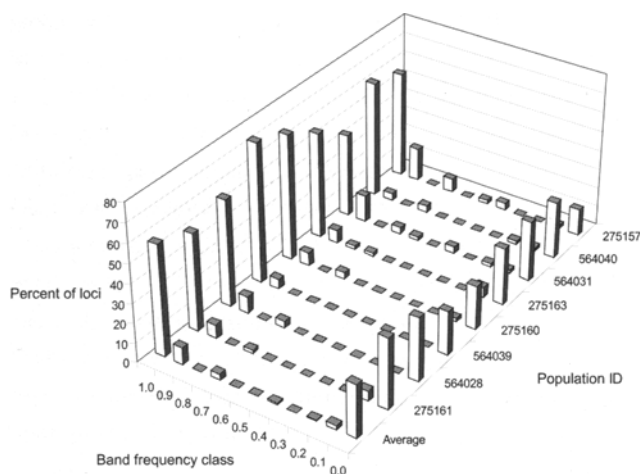
frequency both at 50%, representing RAPD band and blank frequencies, respectively. Two coins flipped in an infinite number of paired trials have a matching frequency (apparent GS) of exactly 50% ( $\frac{1}{4}$  HH +  $\frac{1}{4}$  TT), even though the coins are identical (have an actual GS of 100%). If one composes paired trials of bulk samples (more than one coin), heads is recorded as present if at least one coin comes up heads in the sample. This corresponds to RAPDs in which the band is detected if it is present in at least one plant of the bulk. So, for example, if one flipped two coins at a

time, recorded the results, and then flipped the two coins again, heads would be present three-quarters of the time and absent one-quarter of the time. Thus, the apparent GS of duplicate samples of size 2 is  $(\frac{3}{4})^2$  HH +  $\frac{1}{4})^2$  TT) = 62.5%. This is like bulking the DNA of two plants from a single population into one sample, recording the band status, repeating the process again, and comparing the results. Bulking more plants or coins obviously results in apparent GS closer to the actual 100%, while increased replication does not. Unlike standard coins, band frequencies for polymorphic RAPD loci are not all 50%. But similar calculations can be made for BP value at a given locus and averaged for all loci in a population for any proposed bulk sample size.

A second, similar effect to examine would be that of bulking hypothetical paired samples from *different* populations. There are several assumptions one might use to model such comparisons, but we chose to consider only those loci within a species subject to sample variation, namely, only unfixed loci. When such loci are in common between species (i.e., represent the same band), the observed GS may appear to be less than the actual 100%. Since different populations are being compared, common polymorphic loci may have different BP frequencies. Returning to the coin illustration, the problem now becomes flipping duplicate bulks of two types of coins (with different head frequencies). For example, flipping a single 10% heads coin and then a 90% heads coin would result in matching (GS) of only (9% HH + 9% TT) = 18%. Now a bulk of two coins with 10% heads contains at least one heads 19% of the time, and a bulk of two 90% heads contains at least one

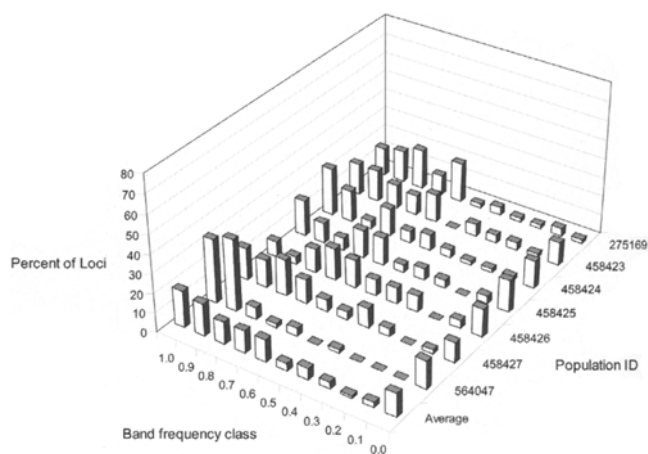


**FIGURE 1.**  
Apparent GS of unfixed loci in common between two hypothetical populations with very different BP frequency distributions.

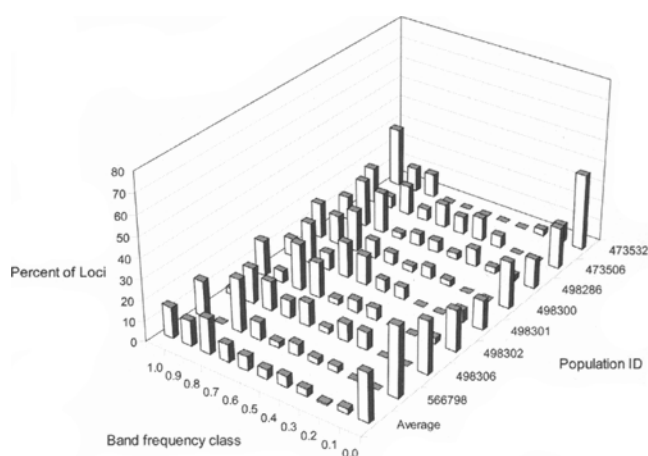


**FIGURE 2.**  
*Solanum fendleri* band frequency distributions.

head 99% of the time. If an infinite number of such two-coin bulk samples were examined in paired trials, the apparent average GS would be the percentage of matches: when heads was present in both samples ( $0.99 \times 0.19$ )  $\approx$  19%, plus when heads was absent from both samples ( $0.01 \times 0.81$ )  $\approx$  1%, for a total of  $\approx$  20%. GS for larger bulk samples can be calculated in the same way. If two populations tended to have mostly extreme BP frequencies (far from 50%) of both types (e.g., either 10% or 90%), three different types of GS have to be averaged, namely, when the loci in common are, by chance, 90% in both populations, 10% in both populations, or 10% in one population and 90% in the other. Figure 1 presents a graph of the expected GS of common polymorphic loci for two hypothetical populations of this kind when compared as paired random bulk



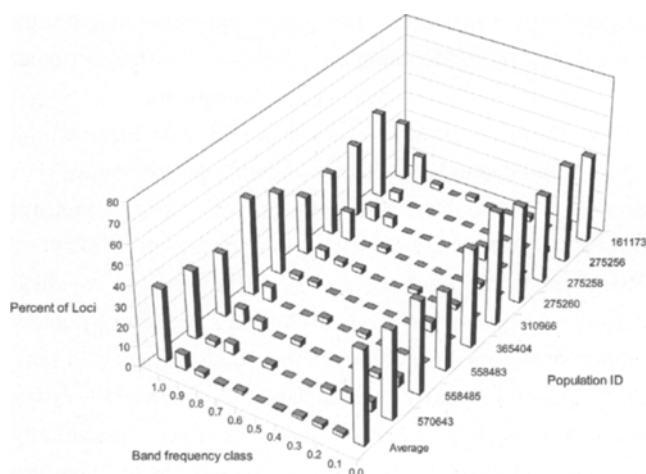
**FIGURE 3.**  
*Solanum jamesii* band frequency distributions.



**FIGURE 4.**  
*Solanum sucrense* band frequency distributions.

samples of various sizes. The graph also includes, for comparison, the expected GS of common loci of two populations that both have all common polymorphic loci at BP frequency of 50%.

The effect of potentially different BP frequency distributions on GS was examined with respect to the empirical species data. To do so, we determined the average BP frequency distribution of unfixed loci for each species (nine classes rounding to 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9), and then calculated the frequency of all possible random pairwise combinations of these BP among species, assuming any such combinations of BP frequencies could represent loci in common between the species. We then calculated the weighted average expected GS of all these loci according to the probability formula explained above.



**FIGURE 5.**  
*Solanum verrucosum* band frequency distributions.

Because the average distribution of BP frequency of loci was similar for inbreds (*fen* and *ver*) and for outcrossers (*jam* and *scr*), the comparisons were combined to present only the expected GS when comparing unfixed loci of selfers to each other, outcrossers to each other, and selfers to outcrossers. These three expected GS levels were graphed against bulk sample size. All of the above calculations were done using Microsoft Excel®.

## RESULTS

The distribution of band frequencies (grouped in classes from 0.0 to 1.0) and the species average are presented in Figures 2-5. Table 1 presents the average (ungrouped) GH for polymorphic bands within each species. *Jam* and *scr* had much higher GH at 29% and 44%, respectively, than *fen* and *ver*. All species differences in GH were highly significant except between *fen* and *ver*, the most homogeneous species with average GH of only 6%-7%. Calculations based on differences in band frequency distributions among species demonstrated that the apparent GS of samples from the same population do not approach 100% unless bulk samples contain 25 to 40 random plants. As expected, larger sample size is most important for promoting adequate parity of duplicate samples of the more heterogeneous species (Figure 6).

Similarly, calculations of the expected apparent GS for hypothetical unfixed loci two species have in common (i.e., with actual GS = 100%) exhibited differences in GS depending on heterogeneity of species compared. Specifically, when such

loci are compared, selfing species compared to each other will appear to be most different, then comparisons of selfers to outcrossers, and finally comparisons of outcrossers to each other. These difference in apparent GS among breeding types is maximum when bulk sample size is around five plants and approaches zero for bulks of 25 to 30 plants (Figure 7).

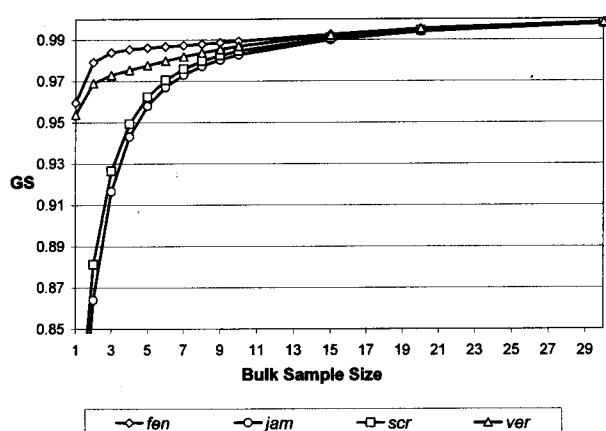
## DISCUSSION

RAPD variation observed was found to be in harmony with expectations based on the breeding systems and ploidies of the species tested. As noted, selfing species, both diploid *ver* and disomic tetraploid *fen*, would be expected to be more uniform, having most of their loci fixed in the banded or bandless state.

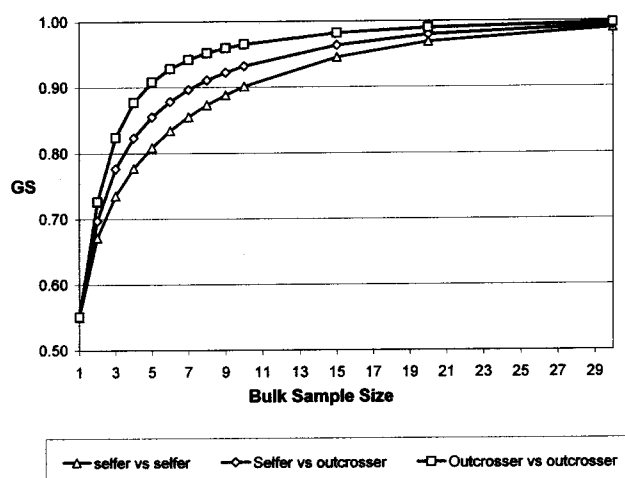
GH represents the estimated proportion of heterozygotes assuming the BP frequencies observed reflect random mating equilibrium, and recognizing that RAPD band "alleles" are detected as though they were genetically dominant. Thus, GH increases with increasing BP frequency until BP is nearly 100% because most of the plants showing the band at such frequencies are expected to be heterozygous. In tetrasomic tetraploids like *scr*, homozygotes are particularly rare, so polymorphic loci make a greater contribution to average GH than they do for diploids. The basis of the GH values given in Table 1 are illustrated in Figures 2-5, where outcrossers *jam* and *scr* have much greater proportions of unfixed loci, particularly ones with high BP frequencies. So high BP frequency at a locus indicates a high proportion of heterozygotes and similar proportion of the band and blank alleles. In contrast, loci with low BP frequencies would represent bands that are at low frequency in the population, i.e., relatively rare. Note from Figures 2-5 that there are relatively few such loci in any of the species tested.

The outcrossing diploid *jam*, at 29.4%, had greater average GH than the selfers. The even higher GH of the tetrasomic tetraploid *scr*, at 44.1%, also fits with the fact that this species carries up to four different genetic alleles per locus. Assume, for example, that RAPDs only detect one polymorphism per locus. There are a maximum of four different bands (and associated blanks) that could be amplified at any given locus for a tetraploid, but only two for a diploid.

All species contained populations that had GH means outside the 95% confidence interval for random samples from a distribution with the species mean (Table 1). This means that populations within species have significant differences in het-



**FIGURE 6.**  
Apparent GS of two random samples from the same population (all loci).



**FIGURE 7.**  
Apparent GS of hypothetical unfixed loci in common between two populations with the BP frequency distributions observed in different species.

erogeneity. GH variation among populations increased with the mean GH for the species. This is to be expected for two reasons. The nature of the binomial distribution is such that random samples from populations with GH approaching zero will be more uniform than those where GH is closer to 50%. Another more mechanistic reason is that inbred populations in nature are expected to be subject to more pressure to homogenize since all recessive alleles are exposed to natural selection. For the same reason, typical germplasm collection at a specific time in the season is more likely to result in a homogeneous sample (e.g., only genetically similar plants will be

fruiting at the same time), and chance unrepresentative sampling during reproduction in the genebank (Widrechner et al. 1989) is more likely to result in homogenizing drift.

As for most crops, *in situ* genetic diversity for potato is much greater than that which we can hope to capture in a genebank with limited resources. So maximizing genebank diversity depends on the best sampling and allocation of effort. These results suggest that breeding system predicts the allele density of populations. Thus, homogeneous inbred populations contain less diversity, and a few plants are likely to sample a large proportion of the population's total diversity. A way to quantify this is to calculate the expected genetic similarity (GS) of random samples from a single population. Figure 6 illustrates how great an impact heterogeneity and small sample numbers can have on RAPD sample variation. The theoretical estimate suggesting that 25 to 30 plants must be bulked to achieve 99.5% similarity of duplicate samples for most potato species has been validated empirically (e.g., Bamberg et al. 2001).

Figure 6 illustrates several useful approximations with respect to sample size. It is apparent that examination of only three or four plants would likely give a poor representation of outcrossers, but a quite adequate one of selfers. If one wanted as much sample parity as practical, it can be seen that with a bulk containing 25 to 30 plants GS is nearly 100% for all species, and adding more plants results in diminished gains in GS. Since RAPD bands appear to be dominant, the generalizations made above based on the detection of one banded plant in a bulk DNA sample parallel, for example, the expectations for finding a single gene or trait in at least one plant of the given sample size.

Collecting strategies may also be influenced by GH. Patterns of diversity in the wild are expected to be more influenced by (and thus correlated with) variation in eco-geographic parameters since all recessive alleles are exposed to selection pressures. Thus, for inbreds, it may be easier to identify locations or environments where further collecting would most likely result in the capture of diversity not already in the genebank (del Rio et al. 2001; del Rio and Bamberg 2004). Also, since selfers contain fewer alleles per population, collection and preservation of more populations of such species may be prudent.

Taxonomic characterization that compares populations and species will also be affected by the interaction of GH and sample size. For example, selfing species (with lower GH)

have about four times as many fixed loci as do outcrossers. These fixed loci contribute to lower sample variation and better resolution of true differences between selfers. But polymorphic loci actually in common between selfing species also make them appear artificially more distinct, especially when small samples are used. This is because selfers tend to have a higher proportion of polymorphic loci that are almost fixed—e.g., not around 50%, but either 10% or 90%. The low (e.g. 10%) BP loci particularly contribute to an erroneous low apparent GS in selfers, as shown in Figure 7 (empirical data) and Figure 1 (hypothetical model). Both of these artificial contributions to breeding system dependent disparities in apparent GS disappear when bulks of 25 to 30 plants are used, at which point GS approaches the true value of 100% for all species (Figure 7).

A germplasm management strategy that assumes more homogeneous inbred populations can be represented by fewer individuals has one caveat. Those few low-frequency segregating alleles would be particularly vulnerable to loss from drift or selection. However, several factors mitigate such losses in selfers. One already mentioned is that potato genebanks that do hand pollinations for seed increase intentionally intermate rather than self plants of all species, thus increasing the partitioning of allelic diversity within, rather than among plants. And, as already noted, the relatively low occurrence of loci with low BP frequency indicates that rare, vulnerable bands are uncommon. Also, maintaining the diversity for particular alleles in particular inbred populations may not be important if both alleles are already fixed in other populations in the genebank, as some empirical evidence has suggested (Bamberg and del Rio 2003).

## ACKNOWLEDGMENTS

The authors wish to express thanks to the University of Wisconsin Peninsular Agricultural Research Station program and staff for their cooperation.

## LITERATURE CITED

- Bamberg, JB, and AH del Rio. 2003. Vulnerability of alleles in the US Potato Genebank extrapolated from RAPDs. *Amer J Potato Res* 80:79-85.
- Bamberg, JB, SD Kiru, and AH del Rio. 2001. Comparison of reputed duplicate populations in the Russian and US potato genebanks using RAPD markers. *Amer J Potato Res* 78:365-369.
- Bamberg JB, MW Martin, JJ Scharfner, and DM Spooner. 1996. Inventory of tuber-bearing *Solanum* species. Catalog of Potato Germplasm-1996. NRSP-6, Sturgeon Bay, WI.
- Bamberg JB, C Singsit, AH del Rio, and EB Radcliffe. 2000. RAPD analysis of genetic diversity in *Solanum* populations to predict the need for fine screening. *Amer J Potato Res* 77:275-277.
- Correll DS. 1962. The Potato and its wild relatives. Texas Research Foundation, Renner, Texas.
- del Rio AH, and JB Bamberg. 2003. The effect of genebank seed increase on the genetics of recently collected potato (*Solanum*) germplasm. *Am J Potato Res* 80:215-218.
- del Rio AH, and JB Bamberg. 2004. Geographical parameters and proximity to related species predict genetic variation in the inbred potato species *Solanum verrucosum* Schlecht. *Crop Sci* 44: 1170-1177.
- del Rio AH, JB Bamberg, and Z Huaman. 1997a. Assessing changes in the genetic diversity of potato genebanks. 1. Effects of seed increase. *Theor Appl Genet* 95:191-198.
- del Rio AH, JB Bamberg, Z Huaman, A Salas, and SE Vega. 1997b. Assessing changes in the genetic diversity of potato genebanks. 2. In situ vs ex situ. *Theor Appl Genet* 95:199-204.
- del Rio AH, JB Bamberg, Z Huaman, A Salas, and SE Vega. 2001. Association of eco-geographical variables with patterns of genetic variation in native wild US potato populations. *Crop Sci* 41:870-878.
- Hawkes JG. 1990. The Potato: Evolution, Biodiversity and Genetic Resources. Belhaven Press, Oxford.
- Hosaka K, and RE Hanneman Jr. 1991. Seed protein variation within accessions of wild and cultivated potato species and inbred *Solanum chacoense*. *Potato Res* 34:419-428.
- Lynch M, and BG Milligan. 1994. Analysis of population genetic structure with RAPD markers. *Mol Ecol* 3:91-99.
- Oliver JL, and JM Zapater. 1984. Allozyme variability and phylogenetic relationships in the cultivated potato (*Solanum tuberosum*) and related species. *Plant Syst Evol* 148:1-18.
- Spooner, DM, DS Douches, and MA Contreras. 1992. Allozyme variation within *Solanum* sect. Petota, ser. Etuberosa (Solanaceae). *Amer J Bot* 79:467-471.
- Widrechner MP, LD Knerr, JE Staub, and K Reitsma. 1989. Biochemical evaluation of germplasm regeneration methods for cucumber, *Cucumis sativa* L. FAO/IBPGR Plant Genet Res Newsletter 88/89:1-4.